Innovative Deep Learning Methods for Computer Aided Drug Design using Big Data

A
Ph.D. Synopsis
Submitted
to

Swami Ramanand TeerthMarathwada University, Nanded

for
Doctor ofPhilosophy
in
Computer Science
under the faculty
of
Science&Technology

Submitted By
Mr. NinadJ.Dani
ResearchStudent

Under Guidanceof
Dr. Nilesh K. Deshmukh
Research Guide&Associate Professor
School of Computational Sciences,
Swami RamanandTeerthMarathwadaUniversity, Nanded
2022-2023

❖ **Table of Contents:**

Title	Page No.
1. Introduction	1
2. Review of Literature	2
3. Objectives of Research	5
4. Research Methodology	5
5. Importance of the study	9
6. Proposed work plan	10
7. References	11

1. Introduction:

This proposal is directed towards the study and development of Novel Deep Learning methods for one of the important aspects in the Computer Aided Drug design lifecycle which is Drug-Target interaction and Molecular Modelling that leverages Big Data. With the exponential growth of available biological and chemical data, the traditional methods for drug discovery and design are facing challenges in efficiently extracting meaningful insights. This proposed study utilizes Deep Leaning techniques and algorithms to analyse vast datasets and will enable accurate predictions for molecular properties, interactions and will find potential drug candidates. Because of the use of Neural networks and advanced data representation techniques, this approach will demonstrate an improved accuracy and efficiency in drug designing tasks. By the use of comprehensive experimentation and validation techniques; the study of the efficacy of proposed Deep Learning method harnessing the big data for advancing the emerging field of Computer Aided Drug Desing will be done.

Computer Aided Drug design, even though can be thought of as a subsidiary in Bioinformatics, is actually becoming and is one of the greatest applications of Computational techniques like Deep Learning for medicinal sciences. The quest for discovering effective drug candidates has been significantly accelerated by the integration of such computational techniques with experimental methods. Computer Aided Drug Design (CADD) has emerged as a pivotal approach for the process of identifying potential drug molecule and their interaction with the biological targets. Currently we are experiencing the proliferation of biological and chemical data at an unprecedented scale and hence the conventional techniques for drug discovery actually faces two challenges: i) The efficient analysis of this abundant data, ii) Extraction of meaningful insights from this oversized data.

This proposal will address these challenges by harnessing the power of Deep Learning in conjunction with Big Data. The novel Deep Learning method capitalizes on the great ability of neural networks to learn the intricate patterns of data, which in turn leads to more accurate predictions and insightful inferences. The primary approach of proposed method is, and not limiting to, the prediction of drug-target interactions and the facilitation of molecular modelling, the two crucial facets in CADD.

The large-scale biological and chemical datasets like ChEMBL, DrugBank, PubChem etc., will be leveraged for the proposed deep learning framework and this aims to unlock the

hidden patterns and relationships between the molecular structures and biological activities. The proposed innovative approach can not only streamline the drug discovery process but also will provide a deeper insight into mechanics of drug-target interactions. The combination of advanced neural network architecture and sophisticated data representation techniques, the proposed approach tries to transcend the limitations of traditional methods and bring efficiency and precision in CADD.

In the subsequent sections we will see the specifics of Literature referred and supported for the proposed study, Specific objectives of proposed study, methodology that can be used and importance with international and significance perspective.

2. Review of Literature:

An interpretable graph-based model for drug-target interaction that makes use of natural language processing categorization was developed by Mehdi and colleagues in their study [1]. In order to find a solution to the DTI issue, they developed a prediction model that makes use of protein binding locations, which incorporates a self-attention mechanism. In this case, the drug target complex is regarded as a sentence together with its biological components, and the sentence classification model from natural language processing is used. The drug and target data are separated and passed to graph attention embedding modules. The converted data is then passed to an attention module combining the attention weights with the concatenated vectors of drug and target data to generate the flatten output layer as the end result that would be used to predict the DIT using NLP techniques. The multi layered perceptron and BiLSTM are used in their algorithm to generate the model. Particularly this paper suggests a Graph-Based Convolutional Neural Network (GCNN) approach for drug target interaction.

A read-cross method was introduced by He and colleagues in their publication [2] for the purpose of predicting the drug target binding affinity via the use of gradient boosting machines. As a result of including the whole interaction spectrum, which ranges from true negative to true positive interactions, the authors of the research provided a technique that can constantly predict the values of binding affinities of the chemicals and proteins. In addition to this, they have presented an additional technique that computes a prediction interval in order to evaluate the degree of confidence in the projected affinity. The most significant obstacle in the process of finding a molecule that preferentially binds to a certain protein is that it is both costly and difficult. In this particular case, more than ninety percent

of the candidate compounds were unsuccessful owing to toxicity problems. This problem is handled in this research by using gradient boosting machines. By passing the Drug Similarity value, Binding Affinity value and Target Similarity value as input to generate the feature vectors and further are passed to the proposed algorithm and the prediction accuracy was noticeable. The datasets used for this are Davis, Metz and KiBA.

One of the innovative approaches that was provided by Ozturk et al. in their study [3], which also demonstrates the ability of Deep Learning for Drug Target interaction prediction, is one of the ground-breaking approaches. In their research, they discovered that the DTI plays a significant role in the process of drug discovery, and that this process can be sped up by using machine learning, and more specifically deep learning. They made use of the datasets from DAVIS and KiBA, and after refining them, they sent them to the layered CNNs in order to construct a learnt model that is capable of predicting the drug-target binding affinity using just sequences of proteins and pharmaceuticals. The use of simplified molecular input line entry system (SMILES) as a one-dimensional representation and the use of one-dimensional convolutional neural networks (CNN) are both discussed in a clear and concise manner. The reader is able to test and experiment with these systems, and the results may be regenerated after careful feature engineering.

One of the uses of this Drug-Target interaction is actually to repurpose the already existing drug and it is explained by Tanoli et. al in their paper [4]. In order to facilitate the interactive investigation of drug-target interaction networks, they have developed a drug target profiler, which is a web-based application that is open-source. Although not being particularly implementing the neural network approach, the paper shows importance of computational methods in drug discovery purposes.

Another computational approach using Big Data for classification strategy is explained in the paper [5] by Ye et. al. An explanation of the drug-target interaction prediction approach that is based on various categorization methodologies is provided in the publication. The targets are divided based on the total number of interactions they have, where one part has smaller number of interactions and another one has comparatively larger interactions. The strategy is tested on five Drug-Target interaction datasets, and it is proven to be competitive with already existing methods and can improve performance with ample amount of data.

Transfer learning and ensemble modelling can also be used for drug target interaction prediction and the same approach was taken by Kao et. al in their paper [6]. The paper

actually introduces a new Deep Learning approach and uses the sequence information of chemical compounds and proteins. And after aggregation of predictions from multiple deep neural networks, gives final output. Using basically most famous DAVIS and KiBA datasets, this approach reduces the chance of overfitting, yields unbiased prediction and achieves good performance.

The convolution on protein sequences is used in the paper [7] by Lee et. al. This paper highlights use of CNN with the Big Data that we can use to predict the Drug-Target interaction. The large-scale DTI information is used for training the model proposed in this paper. The CNN approach was taken actually to reduce the time required for conducting biological DTI experiments, which may take 2~3 years to complete. The researchers focused on pooling maximum convolution results by each filter for each window in the model and that showed the capability of the model to capture the local residue patterns which participate in DTI. And those local features are actually used for regular DTI predictions with other components.

One of the best applications of machine learning in Molecular Modelling is explained and introduced by Wu et. al in their paper [8]. This molecular machine learning is significantly improving because of the increasing availability of larger datasets which helps the algorithms to make more accurate prediction by learning patterns. It uses curated data from ChEMBL, PubChem, ChemSpider and PDB. This paper proposed a work which uses open source package called DeepChem. Using various ensemble machine learning models, it compares their performance and give the reader an idea about the best machine learning approaches for Molecular modelling using large sources of data.

In the paper [9], Buttler et. al states the use of Machine learning for molecular modelling science. The paper enlists the general-purpose machine learning frameworks as well as some machine learning tools for molecules and materials, then further discusses the discovery of new compounds which is quite interesting area as an application of machine learning to discover something new.

When we talk about Big Data in Drug Discovery, the paper [10] by Glicksberg et. al highlights the use of BD in DD. The paper deeply discusses the traditional pre-clinical trial issues and elaborates use of Big Data as a boon for solving such problems. The paper further discusses the tools and software particularly used for effectively using the big data. At the end

the paper outlines a method for using Big Data in drug discovery process to highlight the importance of Big Data.

In an impactful review article by Bhaskaran et. al [11], it primarily focuses on big data generated by patients, hospitals, researchers etc. and how this leads to increase in data for pharmacology, toxicology and pharmaceutics. The problem is to manage and properly use the data. The review article focus on how the researchers can utilize this data to generate meaningful insights. This article also gives us a representation of how the big data generated in clinics, hospitals etc. can be channelled properly.

For molecular designing how the RNN architecture such as LSTM can be used is thoroughly discussed in [12] by Xu et. al. The authors are all from Computer Science background focus on implementing LSTM and GAT to extract global molecular representation by using graph nodes. The paper discusses overall framework to achieve property extraction on Key items highlighting by giving a single molecule as an input. The preprocessing dilemma of using which string when all are representing the same molecule is also discussed and how SALSTM models uses these SMILES strings to solve many such issues is highlighted. The results on eight datasets with either regression or classification model are given and experiments are thoroughly represented.

In the book by Ramsundar et. al [13], the team of authors are extending deep learning for various aspects of deep learning with practical and technical examples in Drug Designing along with many other fields. The book proves to be helpful to inculcate Deep Learning ability for biological sciences in particular for drug discovery. It mentions the life cycle and all aspects also including drug-target interaction and how deep learning is leveraging data to speed up the process.

3. Objectives of Research:

1) To study existing deep learning approaches aiding Computer Aided Drug Design
2) Study Databases and developing data cleaning and refining methods for the Big Data available related to subject
3) Developing own Deep Learning Algorithm for Drug Target Interaction prediction and Molecular Modelling in Computer Aided Drug Design

4) Validate and Evaluate results generated by the algorithm to refine further research

4. Research Methodology

As the deep learning architecture is to be used for the research using big data the steps and methods are as follows:

1) Select Datasets-

   As there are various huge amount of data of molecular components are present, first select the needed sources of data by applying filters.

2) Data Collection and Preprocessing-

   The Big Data available on various databases will be accumulated and pre-processed as to be used as an input to the deep learning model. Preprocessing includes making the data into unified format, removing outliers and duplicate entries wherever needed and cleansing unnecessary parameters.

3) Feature Engineering-

   The received cleaned data, like molecular structures will be converted to appropriate numerical/alphanumerical representations like SMILES, molecular fingerprints or molecular graphs. Then we can extract relevant features from the target molecules or proteins; these methods will also be optimally developed by studying existing methods.

4) Model Selection/Generation-

   The Deep Learning architecture will be selected as a base point and then by refining parameters and hyper-parameters a novel model will be generated that will be efficient and optimal for solving the problem. The model can be a combination of CNN and GNNs if possible and various testing parameters will be utilized for more refinement. Variations of transfer learning will also be used for feature extraction from data.

5) Model Architecture-

   A new architecture for machine learning model will be generated including input layers for drug and target features, hidden layers, activation functions and output layers.

   For molecular modelling, the input layer will take molecular features and then hidden layers, activation functions will be used. In case of the task of molecular generation, Generative Adversarial Networks (GANs) can be used.

6) Training and Validation of Model-

The data split will be done experimentally on the available big data so as to minimize the loss and increasing the accuracy of the model. Loss functions will be selected according to the need as the model will be generated for classification in case of identifying affinity pairs and a new regression model in case of affinity prediction. Molecular modelling will also utilize a version of learning model for generating new molecular structures.

7) Performance Evaluation-

Trained model will be evaluated for its ability to predict the drug target interactions correctly. The metrics like accuracy, recall, F1-score, precision can be used to measure performance. For generative models, we can assess the quality of generated molecular structures and properties using appropriate tools.

8) Model Interpretability and Visualization-

Techniques like attention mechanisms, saliency maps or integrated gradients can be used to interpret model's predictions. For generative models, the output molecule can be visualized by tools that can be integrated with present model language like Python.

9) Hyperparameter Tuning-

Different experiments by changing the hyperparameters, network architectures and optimization strategies will be utilized to enhance the performance of the model. The best hyper parameters can be found by using techniques like grid search.

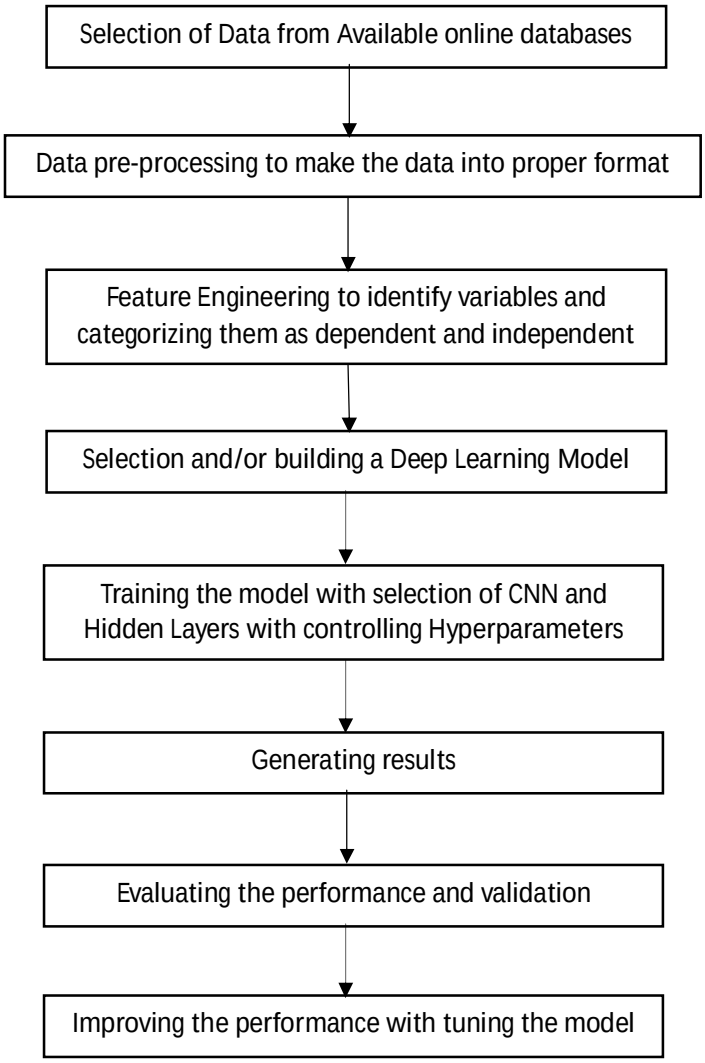
10) Cross-Validation and External Validation-

To check robustness of the model, cross-validation will be carried out on different subsets of data. External validation will be done on individual external data.

11) Continuous Improvement-

Model will regularly be retrained if and when new data arrives, in order to improve predictive and generative capabilities of the model.

The Following Diagram Highlights the Major Steps to carry out the research and Model Building


Selection of Data from Available online databases
Data pre-processing to make the data into proper format
Feature Engineering to identify variables and categorizing them as dependent and independent
Selection and/or building a Deep Learning Model
Training the model with selection of CNN and Hidden Layers with controlling Hyperparameters
Generating results
Evaluating the performance and validation
Improving the performance with tuning the model

5. **Importance of the Study**

I. **International Status:**

The study of Drug Designing using AI and Deep Learning approach has an universal importance. This study will surely contribute in accelerating drug discovery process as a vast amount of Big Data available will be churned and used for predicting interactions and generating new molecular designs. This typically helps in addressing global health challenges. Machine learning enables identification of patient specific treatments based on genetic and molecular data related with the patient hence proper personalized medicines can be given. In case of molecular design, the study helps not only pharmaceutical industries but also material science studies.

The study makes sure to generate computer models which cuts down the requirements of using laboratories for doing compound experimentations hence optimizing cost benefits.

Repurposing of existing drugs has been a fundamental goal in drug designing and the proposed study helps in speeding up identifying such drugs which can be repurposed for other diseases.

Such Deep Learning experimentations are also being done globally in international universities and hence can generate future collaborative opportunities with international universities.

II. **Significance of the Study:**

Using Deep Learning with Big Data for Drug Target Interaction prediction and Molecular modelling has several significant factors as it represents a paradigm shift in the field of Drug Design because of addition of Computer Aided ability to the field:

- Enhanced Accuracy and Prediction because deep learning models can capture and identify complex relationships even from the big data that the study will be targeting. The possibility of designing novel drug molecule also increases because of this.
- Novel Interaction generation ability because of deep learning makes sure the unbiased exploration of vast chemical and biological data provided to the model.
- Leveraging the big data which is present in the diverse formats like biological, chemical and omics data which when integrated and analysed

by Deep learning can provide unexplored insights about the data. Which may also lead to unexplored molecular structure design.

- Deep learning also helps in predicting potential off-target interactions which traditional processes may have always missed, and this helps improving drug safety profile and minimizing adverse side effects.
- One of the important aspects of this study is the opportunity to have a cross disciplinary collaboration between computer scientists and biologists, chemists and pharmacologists.
- The power of deep learning has always been data driven decision making and that same power is now being provided to the field of drug design with the help of Big Data.

6. **Proposed work Plan:**

The proposed work is to be implanted using Deep Learning model building using Tools like Jupyter Notebook (Anacoda), Google Colab Notebook and Modeller tool.

I. 2022-2023:

- Collection of research related papers, books and doing review of literature
- Completing Course-Work
- Study of Deep Learning and Bio-Informatics

II. 2023-2024:

- Databases study and selection / Collection
- Pre-Processing of Data
  - Making in appropriate format
  - Clansing of outliers
- Selection of Deep Learning Base Algorithm for DTI
- Selection of Deep Learning and Graph Adversarial Neural Network for Molecular modelling algorithm
- One or two research papers publication

III. 2024-2025:

- Deep Learning model building with selected and customized algorithm

- Experimenting with different datasets for prediction accuracy and adjusting layers and hyperparameters in the algorithm accordingly
- Refining Molecular modelling techniques according to optimized results
- One or two research papers publications

7. **References:**

[1] Mehdi Yazdani-Jahromi et. al, "AttentionSiteDTI: an interpretable graph-based model for drug-target interaction prediction using NLP sentence-level relation classification", July 2022
https://doi.org/10.1093/bib/bbac272

[2] Tong He et. al, "SimBoost: a read-across approach for predicting drug–target binding affinities using gradient boosting machines", 2017
https://doi.org/10.1186/s13321-017-0209-z

[3] Ozturk et. al, "DeepDTA: deep drug–target binding affinity prediction", 2018
https://doi.org/10.1093/bioinformatics/bty593

[4] Tanoli et. al, "Interactive visual analysis of drug–target interaction networks using Drug Target Profiler, with applications to precision medicine and drug repurposing", 2020
https://doi.org/10.1093/bib/bby119

[5] Ye et. al, "Drug–target interaction prediction via multiple classification strategies", January 2021

[6] Kao et. al, "Toward Robust Drug-Target Interaction Prediction via Ensemble Modelling and Transfer Learning", July 2021

[7] Lee et. al, "DeepConv-DTI: Prediction of drug-target interactions via deep learning with convolution on protein sequences", May 2019
https://doi.org/10.1371/journal.pcbi.1007129

[8] Wu et. al, "MoleculeNet: a benchmark for molecular machine learning", Oct. 2017
https://doi.org/10.1039/c7sc02664a

[9] Buttler et. al, "Machine learning for molecular and materials science", July, 2018
https://doi.org/10.1038/s41586-018-0337-2

[10] Glicksberg et. al, "Leveraging Big Data to Transform Drug Discovery", 2019

[11] Bhaskaran et. al, “A Survey on Big Data in Pharmacology, Toxicology and Pharmaceutics”, Dec. 2022
[12] Xu et. al, “Molecular Property Prediction by Combining LSTM and GAT”, Mar, 2023
[13] Ramsundar et. al, “Deep Learning in the Life Sciences”, 2019

Mr. Ninad J. Dani
Research Student